CLAIMS

What is claimed is:

1. A system for allocating a resource to a service request representing a request for a category of service selected from amongst a plurality of possible categories, comprising:

first logic for selecting, responsive to the selected category, a policy from amongst a plurality of possible policies; and

second logic for applying the selected policy to allocate a resource to the request selected from one or more candidate resources.

2. The system of claim 1 wherein the selected policy is a load balancing policy.

3. The system of claim 1 wherein the selected category of service is a content-enabled category.

4. The system of claim 1 wherein the selected category of service is a content-independent category.

5. The system of claim 1 implemented in hardware as one or more finite state machines.

6. The system of claim 1 wherein the first logic is configured to determine the policy through an access to a table using an index derived from the service request.

7. A system for allocating a resource to a service request comprising:

first logic for determining one or more candidate resources using a hierarchical arrangement of data structures, the hierarchical arrangement having a plurality of levels; and

second logic for selecting one of the candidate resources, and allocating the selected resource to the service request.

8. The system of claim 7 wherein each of the data structures is a table.

9. The system of claim 8 wherein an index derived from an entry in a table at one level of the hierarchical arrangement is used to obtain an entry in the table at a next level of the hierarchical arrangement.

10. The system of claim 7 wherein the resource is a server, and the hierarchical arrangement comprises a service index table, a super-group table, and a server group table.

11. The system of claim 10 wherein an index to the service index table is derived from the service request, and the index is used to access an entry in the service index table specifying a super-group to be allocated to the request, and a load balancing policy.

12. The system of claim 11 wherein an index derived from the super-group allocated to the request is used to access an entry in the super-group table specifying one or more server groups which are candidates for allocating to the request.

13. The system of claim 12 wherein one of the candidate server groups is allocated to the request through application of a suitable policy.

14. The system of claim 13 wherein an index derived from the server group allocated to the request is used to access an entry in the super-group table specifying one or more servers which are candidates for allocating to the request.

15. The system of claim 14 wherein one of the candidate servers is allocated to the request through application of the load balancing policy specified by the entry in the service index table.

16. A system for allocating a resource to a service request comprising:

first logic for specifying a plurality of resources which are candidates for allocating to the request; and

second logic for accessing in parallel loading information for each of the candidate resources; and

third logic for allocating one of the candidate resources to the request responsive to the accessed loading information.

17. The system of claim 16 wherein the loading information for each of the candidate resources is replicated across a plurality of memories, and the second logic is configured to access each of the memories in parallel to obtain the loading information.

18. The system of claim 16 wherein the third logic is configured to allocate one of the candidate resources to the request through application of a load balancing policy to the loading information for the candidate resources.

19. A system for allocating a resource to a service request comprising:

first means for determining one or more candidate resources using a hierarchical arrangement of data structures, the hierarchical arrangement having a plurality of levels; and

second means for selecting one of the candidate resources, and allocating the selected resource to the service request.

20. A method of allocating a resource to a service request representing a request for a category of service selected from amongst a plurality of possible categories, comprising:

determining a policy responsive to the selected category; and

applying the policy to allocate a resource to the request selected from one or more candidate resources.

21. The method of claim 20 wherein the policy is a load balancing policy.

22. The method of claim 20 wherein the policy is determined through an access to a table using an index derived from the selected category of service.

23. A method of allocating a resource to a service request comprising:

determining one or more candidate resources using a hierarchical arrangement of data structures, the hierarchical arrangement having a plurality of levels; and

selecting one of the candidate resources, and allocating the selected resource to the service request.

24. The method of claim 23 wherein each of the data structures in the hierarchical arrangement is a table.

25. The method of claim 24 further comprising deriving an index from an entry in a table at one level of the hierarchy, and using the index to access an entry in a table at a next level of the hierarchy.

26. The method of claim 23 wherein the determining step comprises:

deriving an index to a service index table from the service request;

using the index to access an entry in the service index table; and

allocating a super-group to the request and determining a load balancing policy responsive to the entry in the service index table.

27. The method of claim 26 wherein the resource is a server, and the determining step further comprises:

deriving an index to a super-group table from the super-group allocated to the request;

using the index to access an entry in the super-group table;

determining from the entry one or more server groups which are candidates for allocating to the request; and

allocating one of the candidate server groups to the request.

28. The method of claim 27 further comprising allocating one of the candidate server groups to the request through application of a suitable load balancing policy.

29. The method of claim 27 wherein the determining step further comprises:

deriving an index to a server group table from the server group allocated to the request;

using the index to access an entry in a server group table; and

determining from the entry the one or more servers which are candidates for allocating to the request.

30. The method of claim 23 wherein the resource is a server, further comprising allocating the selected server to the request only if a server is not allocated to the request through application of a persistence policy.

31. A method of allocating a resource to a service request comprising:

a step for determining one or more candidate resources using a hierarchical arrangement of data structures, the hierarchical arrangement having a plurality of levels; and a step for selecting one of the candidate resources, and allocating the selected resource to the service request.

32. A method of allocating a resource to a service request comprising:

specifying a plurality of resources which are candidates for allocating to the request; and

accessing in parallel loading information for each of the candidate resources; and

allocating one of the candidate resources to the request responsive to the accessed loading information.

33. The method of claim 32 wherein the resource is a server, and the accessing step comprises accessing in parallel the loading information from a server loading table replicated across a plurality of memories which are accessible in parallel.

34. The method of claim 32 wherein the allocating step comprises allocating one of the candidate resources to the request responsive to application of a load balancing policy to the accessed loading information.

35. The system of claim 1 wherein the resource is a server.

36. The system of any of claims 7 or 19 wherein the one or more candidate resources are servers.

37. The method of claim 20 wherein the resource is a server.

38. The method of any of claims 23 or 31 wherein the one or more candidate resources are servers.

39. The method of claim 32 wherein the plurality of resources are servers.

40. The system of any of claims 1, 7, 16, or 19, wherein the resources are servers.

41. The system of any of claims 1, 7, 16, or 19, wherein the service requests are in the form of or spawned by packets.

42. The system of any of claims 1, 7, 16, or 19, implemented as one or more engines.

43. The method of any of claims 20, 23, 31, and 32, wherein the resources are servers.

44. The method of any of claims 20, 23, 31, and 32, wherein the service requests are in the form of or spawned by packets.

10074462.052402